Text Classification: The Case of Multiple Labels

Victoria Bobicev
Applied Informatics Department, Technical University of Moldova,
Computers, Informatics and Microelectronics faculty,
Chișinău, Moldova
v_bobicev@mail.utm.md

***Abstract*—Analysis of subjectivity is the actively developed direction of research in text mining. The paper presents machine learning experiments on classification of sentiments in forum texts. We explore the difficult task of classification when texts are labeled by several sentiment labels and in this condition we reach the average F-measure equal to 0.805.**

Keywords–machine learning; natural language processing; text classification; sentiment analysis; multi-label classification.

I. Introduction

During the last decade the Internet has been increasingly used by people not just as users but as producers of online information. A study of user generated content (UGC) from 2008 [1] claimed that 35% of the USA Internet users have contributed at least once a UGC to the Web and the same tendency was observed in Europe, Japan and Korea.

The recent survey [2] emphasizes that emotions are essential part of human interaction. This implies that analysis of text sense would be incomplete without its emotional component. User generated texts reflects the writer's emotional state which can be identified by using various methods. A growing amount of research papers describe various studies of sentiment detection, emotion recognition and opinion classification in user-generated context.

Extraction and analysis of sentiments, opinions, attitudes, emotions, perceptions and intentions is one of the most asked-for types of text analysis, as was pointed out in Seth Grimes' Text Analytics Report 2014[1].

This paper analysis the problem of sentiment detection in health-related forum posts.

II. Related Work

The sub-field of natural language processing named "Sentiment Analysis" explores various methods and algorithms for detection of such subjective phenomena such as opinions, sentiments, emotions and writer's mood and creates numerous applications for these purposes [3].

Health-related forums are subjects of increasing number of studies. Health-related sentiments were studied in [4]. Around 600 sentences from the forum messages about hearing problems were annotated with positive, and negative labels and used in machine learning experiments which achieved F1 measure up to 0.685. Changes of sentiments in a health-related online community were studied in [5]. The forum messages were classified into positive and negative using message texts. The authors demonstrated that the initial negative posts were often followed by positive posts of the same participant. In [6] a medical domain lexicon was built on the base of user reviews on drugs with ratings from negative to positive. They achieved F-score of 0.62 for the positive class, 0.48 for the negative class and 0.09 for the neutral class. We used more sophisticated set of sentiment labels which better reflected forum participants' interaction.

In [7] several machine learning algorithms were applied for classification of the data set we used in our work. The best achieved F-measure was 0.518. [8] continued the work on the same data experimenting with various lexicons and machine learning algorithms; the results reported for each label separately. The best f-measure for *encouragement* was 0.67. Our work differs from this one in using multiple labels for each message which allowed us to achieve better classification results.

III. Data and Task Description

We worked with the data collected from the IVF forums (ivf.ca) and manually annotated using several labels. The data and annotation were described in [7] and [8].

We obtained the data where each forum post was considered as an annotation unit and annotated with up to three labels. Thus, our aim was to detect all of these labels for every analyzed post.

A. Data Description

The IVF forums consist of threads each of them presenting a conversation on a topic indicated by the thread name. We worked with 80 threads which contained 1321 posts in total. The average length of a thread was 16-17 posts. Each post was annotated manually by several annotators [8]. After several rounds of annotations the number of labels for each post ranged from one to three.

The themes of discussions on the forums were connected to health and fertility problems, fertility treatment and in vitro fertilization. Health, pregnancy and babies connected issues evoke very strong emotions and sentiments which are not easy classified in positive and negative only. Thus, posts in these discussions were annotated with specific labels: *confusion*, *encouragement*, *gratitude* and *facts*. The first three labels indicated sentiments addressed by the post authors to the other forum participants. The fourth labels indicated emotion free narrative.

[1] http://altaplana.com/grimes.html

978-1-4673-8197-0/16/$31.00 ©2016 IEEE

As the annotated posts had the length of 128 words on average and contained several sentences different parts of them expressed different sentiments and many posts contained some neutral factual information. This was the cause why they were labeled with multiple labels. The annotation statistics was the following: 658 posts were annotated with only one label; 605 posts had two labels; 58 posts had three different labels. The most frequent label was *facts*, 954 posts were annotated with *facts,* 642 posts were annotated with *encouragement*, *confusion* appeared in 285 posts, and *gratitude* in 161 posts.

B. The Task and the Instruments

Our current experiments had several objectives. The main goal was to handle multiple annotations of posts; the second one was to compare several sentiment lexicons and find a better one for these particular texts.

There are two main approaches to multi-label classification problems: a) *problem transformation*, and b) *algorithm adaptation* [9]. We used the first method, namely, we analyzed every label apart for the post in question. This means that we run one experiment for each label detecting whether each post was marked by this label or not. In our case this method can be applied as in most texts different fragments express different sentiments. For example, the post can be started by complains and description of the authors problem and worry, then it could be the part where the author story is described in more o less objective narrative. The final part usually contain some warm words addressed the other forum participants. Thus, the sentiments are mostly independent in text and can be detected individually.

We used Machine learning toolkit WEKA[2] to experiment with our data. WEKA [10] is open source software issued under the GNU General Public License and contains a collection of machine learning algorithms for data mining tasks such as a Bayesian algorithms, decision trees, Support Vector Machine, Instance-Based learner, Logistic Regression, etc.

C. Resources Used in Experiments

One of the most important elements in machine learning is the feature set. In most sentiment analysis tasks this set was based on sentiment lexicon. We collected nine lexicons and performed experiments with every one of them. Below is the list of used lexicons with the short description.

1. (SWN) SentiWordNet lexicon assigns each synset of WordNet with three sentiment scores: positivity, negativity, objectivity [11]. We selected synsets with non-zero score for positivity or negativity.
2. (SS) SentiStrength assigns a score from 1(no positivity) to 5 (extremely positive), and -1(no negativity) to -5 (extremely negative) [12]. We selected the terms with the score <-2 and >2.
3. (DM) DepecheMood contained word - emotion matrix, where the list of emotions was CONCERNED, AFRAID, AMUSED, ANGRY, ANNOYED, DONT_CARE, HAPPY, INSPIRED, and SAD [13].
4. (HA) HealthAffect was the domain-specific lexicon created specifically for these texts taking into consideration the specific set of sentiment labels used for annotation [8].
5. AFINN was based on Affective Norms for English Words (ANEW) lexicon and ranks words on a scale from strongly negative to strongly positive [14].
6. (GE) General Enquirer assigned words into positive and negative sentiment, and mood categories [15]. We selected the words with positive and negative sentiment labels only.
7. Hashtag Affirmative and Negated Context Sentiment Lexicon was created on the base of tweet words which had sentiment score from -10 to 10 [16]. We used the terms with scores either < -2 or > 2.
8. Sentiment140 Lexicon was also created from tweets [16] and had the similar structure with the Hashtag Affirmative and Negated Context Sentiment Lexicon.
9. NRC Word-Emotion Association Lexicon assigned to each word scores for eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive) [17]. We used words with non-zero scores for negative or positive sentiment.

We also experimented with eight lexicon combinations listed below. The first four combinations presented intersections of the lexicons and the last four were unions of the same lexicons.

10. Intersection of Hashtag Affirmative and Negated Context Sentiment Lexicon, Sentiment140 and NRC Word-Emotion Association Lexicon.
11. Intersection of Hashtag Affirmative and Negated Context Sentiment Lexicon, Sentiment140, NRC Word-Emotion Association Lexicon, AFINN and General Enquirer.
12. Intersection of SentiWordNet, SentiStrength, and DepecheMood lexicons.
13. Intersection of SentiWordNet, SentiStrength, DepecheMood and HealthAffect lexicons.
14. Union of Hashtag Affirmative and Negated Context Sentiment Lexicon, Sentiment140 and NRC Word-Emotion Association Lexicon.
15. Union of Intersection of Hashtag Affirmative and Negated Context Sentiment Lexicon, Sentiment140, NRC Word-Emotion Association Lexicon, AFINN and General Enquirer.
16. Union of SentiWordNet, SentiStrength, and DepecheMood lexicons.
17. Union of SentiWordNet, SentiStrength, DepecheMood and HealthAffect lexicons.

IV. Experiments and Results

A. Experiment Organisation

We created 17 feature sets comparing sources described and numerated above with the words of our texts and selecting from the lexicons only terms which appeared in our texts. Each of these sets was used for detecting each of the four labels apart.

Naïve Bayes (NB) as the simplest classifier was used as a baseline.

[2] http://www.cs.waikato.ac.nz/ml/weka/

B. Obtained Results

We experimented using Bayesian algorithms (DMNBtext, NBMultinomial) and Support Vector Machine (SVM, SMO realization) provided by WEKA. We present only the tables with the best results for each label and for average value. All figures presented in the tables and figures are macro-averaged F-measures obtained in 10-fold cross-validation experiments. F-measure allows direct comparison with the previous works on the same data set.

Table I
THE BEST F-MEASURE FOR *CONFUSION* WAS OBTAINED USING UNION OF SS SWN DM HA LEXICONS AND NBMULTINOMIAL ALGORITHM

algorithms labels	**NB**	**DMNBtext**	**NBMultinomial**	**SMO (SVM)**
confusion	0.714	0.802	**0.816**	0.786
encouragement	0.594	0.722	0.727	0.692
gratitude	0.701	0.889	0.865	0.876
factual	0.756	0.808	0.805	0.794
Average	0.691	**0.805**	0.803	0.787

Table III
THE BEST F-MEASURE FOR *ENCOURAGEMENT* AND *GRATITUDE* WAS OBTAINED USING HEALTHAFFECT LEXICON AND DMNBTEXT ALGORITHM

algorithms labels	**NB**	**DMNBtext**	**NBMultinomial**	**SMO (SVM)**
confusion	0.732	0.787	0.791	0.758
encouragement	0.640	**0.736**	0.731	0.683
gratitude	0.876	**0.899**	0.898	0.869
factual	0.768	0.726	0.741	0.745
Average	0.754	0.787	0.790	0.764

Table IIII
THE BEST RESULT FOR *FACTUAL* WAS OBTAINED USING UNION OF OF SS SWN DM LEXICON AND DMNBTEXT ALGORITHM

algorithms labels	**NB**	**DMNBtext**	**NBMultinomial**	**SMO (SVM)**
confusion	0.714	0.794	0.796	0.782
encouragement	0.585	0.700	0.690	0.677
gratitude	0.669	0.866	0.846	0.86
factual	0.743	**0.809**	0.759	0.805
Average	0.678	0.792	0.773	0.781

The best results for each label and for averaged F-measure are in bold and underlined.

C. Comparison of Lexicons

Fig. 1-5 present the comparisons of 9 lexicons (numbered 1-9 on the axe X), 4 lexicon intersections (numbers 10-13) and lexicon unions (numbers 14-17 on the axe X). F-measure is presented on the axe Y.

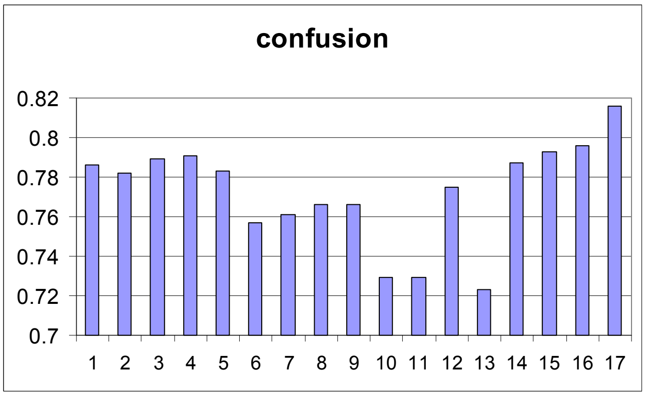

Fig. 1. F-measure obtained for the label *confusion* using 9 lexicons and 8 combinations of these lexicons (see the numeration in part III C)

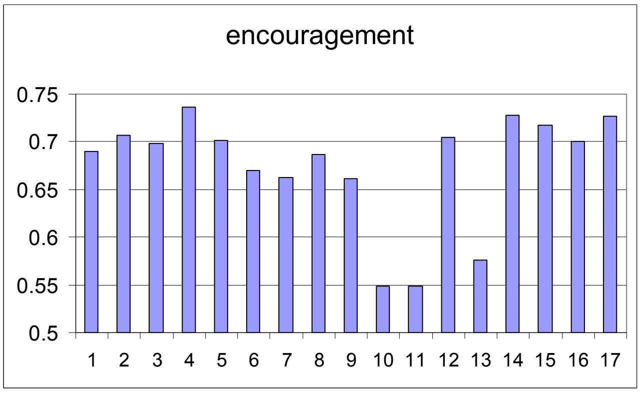

Fig. 2. F-measure obtained for the label *encouragement* using 9 lexicons and 8 combinations of these lexicons (see the numeration in part III C)

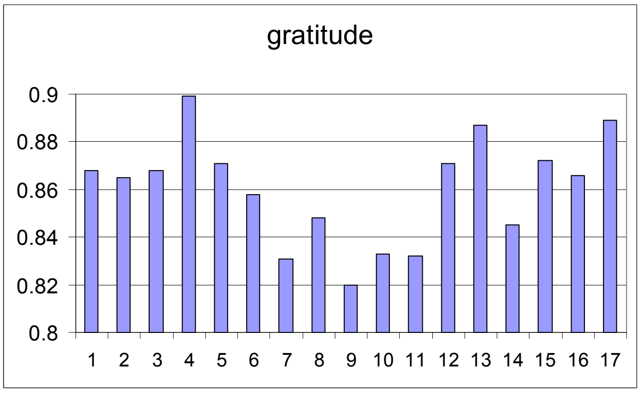

Fig. 3. F-measure obtained for the label *gratitude* using 9 lexicons and 8 combinations of these lexicons (see the numeration in part III C)

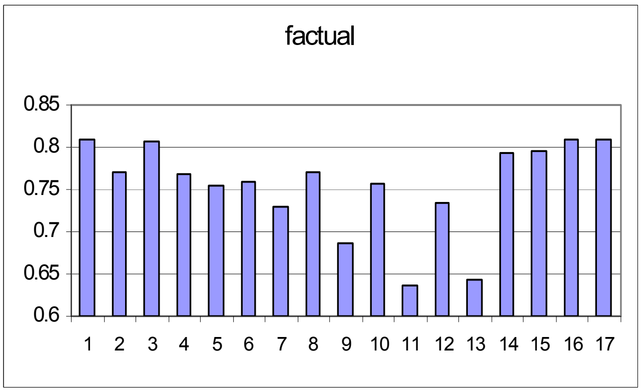

Fig. 4. F-measure obtained for the label *factual* using 9 lexicons and 8 combinations of these lexicons (see the numeration in part III C)

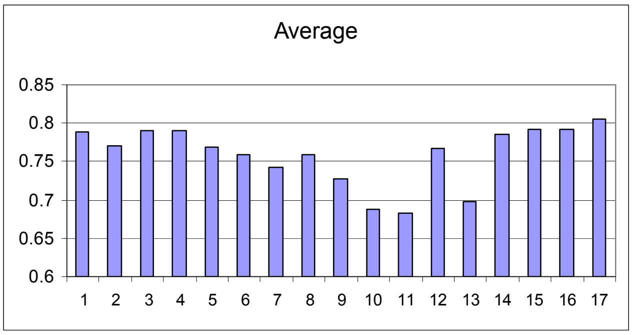

Fig. 5. The average F-measure obtained for all four labels using 9 lexicons and 8 combinations of these lexicons (see the numeration in part III C)

V. Discussion and Conclusions

As it is seen on the diagrams, we cannot definitely name the best lexicon. Surprisingly, HealthAffect created especially for this data set was not the better one in all cases, for example, in detection of the label *facts*. Intersections of the lexicons gave the worst results but unions performed better than individual lexicons in most cases. Obviously, unions of lexicons were the largest feature sets which slowed the classification process. The most surprising result was obtained for the label *facts*; the union of sentiment lexicons was better without HealthAffect despite the fact that it contained a special set of words created for detection of this label. We plan to analyze all our sets of features more detailed in order to select the best set of features for our task.

In general, the results were quite good, the best result for confusion was 0.816 and for gratitude 0.899. Even the label factual recognition reached F-measure 0.809. The only category that was recognized worse than 0.8 was encouragement. It happened probably because this attitude was expressed in more specific way, not with special words but with more complex expressions. These expressions were not as easy to detect using just word - based features. Nevertheless, the average result was 0.805 which is rather good one for sentiment recognition tasks.

References

[1] X. Ochoa, E. Duval, “Quantitative Analysis of User-Generated Content on the Web,” First International Workshop on Understanding Web Evolution (WebEvolve2008), China 2008.

[2] D. Sahni, G. Aggarwal, “Recognizing Emotions and Sentiments in Text: A Survey,” International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 5, 2015.

[3] M. Cieliebak, O. Dürr, and F. Uzdilli, “Potential and Limitations of Commercial Sentiment Detection Tools,” ESSEM 2013 Emotion and Sentiment in Social and Expressive Media, Italy, 2013.

[4] Ali, T., D. Schramm, M. Sokolova, D. Inkpen: Can I Hear You? Sentiment Analysis on Medical Forums. IJCNLP 2013: 667-673.

[5] Qiu, Baojun, et al. "Get online support, feel better--sentiment analysis and dynamics in an online cancer survivor community." Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing.

[6] Goeuriot, L., Na, J. C., Min Kyaing, W. Y., Khoo, C., Chang, Y. K., Theng, Y. L., & Kim, J. J. (2012, January). Sentiment lexicons for health-related opinion mining.In Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium (pp. 219-226).ACM

[7] M. Sokolova, & V. Bobicev, “What Sentiments Can Be Found in Medical Forums?,” RANLP 2013, Bulgaria, 2013.

[8] V. Bobicev, M. Sokolova, M. Oakes, “What Goes Around Comes Around: Learning Sentiments in Online Medical Forums”, Journal of Cognitive Computation, 2015.

[9] G. Tsoumakas, I. Katakis, “Multi Label Classification: An Overview”, International Journal of Data Warehousing and Mining, David Taniar (Ed.), Idea Group Publishing, 3(3), pp. 1-13, 2007.

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, Ian H. Witten, “The WEKA Data Mining Software: An Update,” SIGKDD Explorations, Volume 11, Issue 1, 2009.

[11] S. Baccianella, A. Esuli, and F. Sebastiani. “Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining,” Proceedings of the 7th LREC, 2010.

[12] M. Thelwall, K. Buckley, & G. Paltoglou, “Sentiment strength detection for the social Web,” Journal of the American Society for Information Science and Technology, 63(1), p. 163-173, 2012.

[13] J. Staiano, and M. Guerini, “DepecheMood: a Lexicon for Emotion Analysis from Crowd-Annotated News,” Proceedings of ACL-2014.

[14] F. Nielsen, “A new ANEW: evaluation of a word list for sentiment analysis in microblogs,” Proceedings of the ESWC2011 Workshop on ’Making Sense of Microposts’: Big things come in small packages, 93–98, 2011.

[15] P. Stone, D. Dunphy, M. Smith, and D. Ogilvie, “The General Enquirer: A Computer Approach to Content Analysis,” MIT Press, Cambridge, MA, 1966.

[16] S. Mohammad, S. Kiritchenko, and X. Zhu, “Building the state-of-the-art in sentiment analysis of tweets,” Proceedings of the 7th international workshop on Semantic Evaluation Exercises (SemEval-2013), 2013.

[17] S. Mohammad, and P. Turney, “Crowdsourcing a Word-Emotion Association Lexicon,” Computational Intelligence, 29 (3), 436-465, 2013..